

Lec 11 classifiers

* classification : assign labels to objects based on object's attributes.

examples \rightarrow Naïve Bayesian & decision trees.

Naïve Bayesian

\rightarrow Probabilistic classifier based on Bayes' law and naïve conditional independent assumptions.

Input variables are discrete with variations to ~~variables~~ are algorithms that work with continuous variables as well.

output

* probability score : Proportional to true probability

* class label : based on highest probability score.

Bayesian - use cases

* Preferred for many text classification problems.

- * use cases

\rightarrow Spam Filtering

\rightarrow Fraud detection.

في مثال مطلق و (classifier) هينز فيا بعد
 3 ملف و (examples)

* Bayes' Law

$$P(c|A) = \frac{P(A|c) P(c)}{P(A)}$$

Naïve Bayesian classifier

Reasons to choose	Reasons Cautions.
Handles missing values quite well	Numeric variables have to be discrete intervals.
Robust to irrelevant variables	Sensitive to correlated variables "Double-counting"
Easy to implement	Not good for estimating probabilities stick to class labels
Easy to predict data	
Resistant to over-fitting	

* What gives the naïve Bayesian the advantage of being computationally inexpensive?

→ It is computationally efficient.

↳ handles very high dimensional problems.

↳ " categorical variables with lot of levels.

Lec 12

Decision tree classifier

input variables can be continuous or discrete

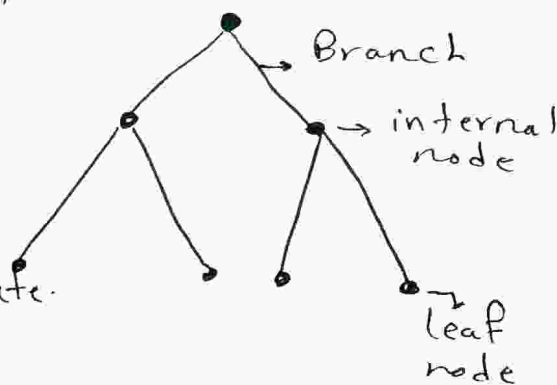
output

↳ tree describes the decision flow.

↳ trees can be converted to set of decision rules.

Branches → outcome of decision

internal node → the decision or test points. each refers to single variables or attribute.



* Decision tree classifier use cases

1] When if-then conditions are preferred to linear models

- ↳ a) Customer segmentation.
- ↳ b) Financial decisions.
- ↳ c) Fraud detection.

2] When series of questions (Yes/No) are answered to arrive at classification.

↳ Biological species classification.

3] مثال من (classifier) في الحياة اليومية
مثال من (لا في من فاهي اوي)

* weak learners

- ↳ It is components (short decision trees)
- ↳ set of predictive models which will all vote and we take decisions based on combination of votes (used in ensemble techniques)

Decision tree classifier

Reasons to choose	Cautions.
a) Easy to Score data.	a) decision Surfaces Can only be axis-aligned.
b) Computationally efficient to build.	b) related to over-fit Problem.
c) handles variables that have non-linear effect on outcome.	c) in practice, decision rules can be complex.
d) decision rules are easy to understand. (in principle)	d) deep tree structure is sensitive to small changes in training data.

Typical questions	recommended methods.
* Do I want class Probabilities rather than class labels.	Logistic regression Decision tree
* Do i want insight into how variables affect model?	Logistic & Decision tree
Is Problem high-dimensional?	Naïve Bayes
Do I suspect some of inputs are correlated?	Decision & Logistic
Do I suspect some of inputs are irrelevant?	Decision tree & Naïve Bayes
Are there mixed variables types	Decision tree & Logistic regression.

Lec 13

Time series

↳ basic research methodology in which data for one or more variables collected for many observations at different time periods.

Key component of time-series

1] Trend Component

↳ trend is long term movement in a time-series.
Underlying direction that can be +ve or -ve.

2] Seasonal Component

↳ component of variation in time series which is dependent on time of year.

~~describes regular variation~~

3] Cyclic Component

↳ cyclical variations of non-seasonal nature, whose periodicity is unknown.

4] Random Component

↳ Random or chaotic values left over when other components of series have been accounted for.

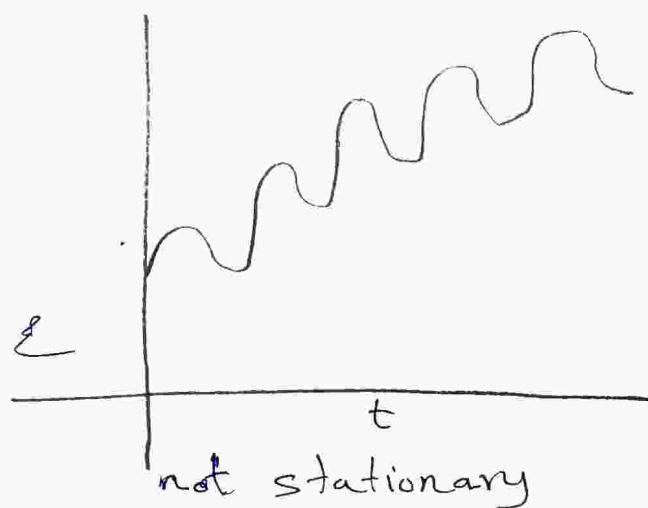
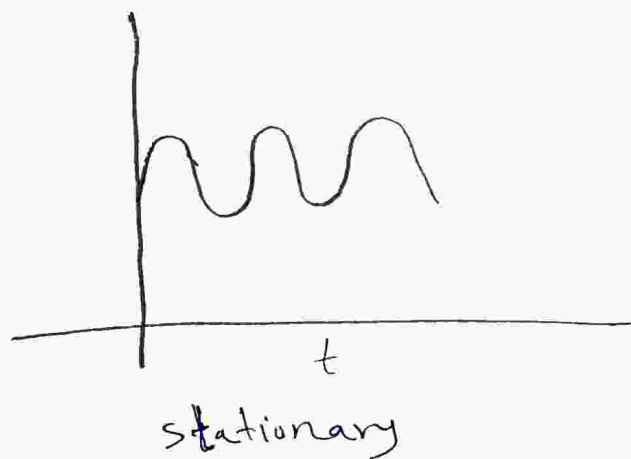
* use cases of time series

- 1) Economics / Finance (profits, imports)
- 2) Environment (amount of pollutants)
- 3) Medicine: Blood pressure measurements.
- 4) Sociology: Crime rate

* stationary sequences:

↳ random sequence in which joint probability distribution does not vary over time.

↳ mean, variance and autocorrelations don't change in sequence over time.



* Constant mean & Variance

↳ mean of series shouldn't be function of time.

↳ variance " " " " " " " " " " " "

Notes

- variance refers to spread of data set
- Covariance " " measure of how two random variables will change together.
- used to calculate correlation between variables.

Detrending

↳ Pre-processing step to prepare time series for analysis by methods that assume stationarity.

سلسلة (non-stationary) → السلسلة
(de-trend) السلسلة

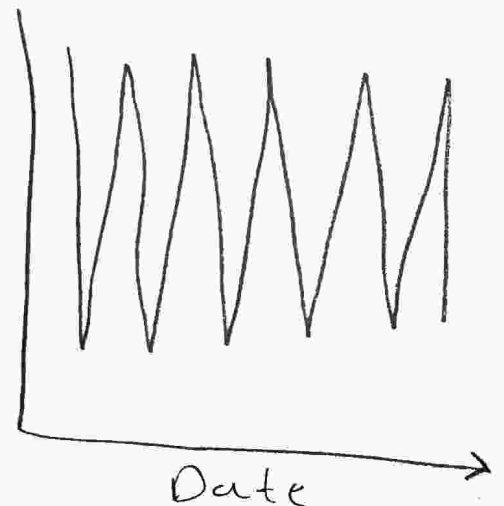
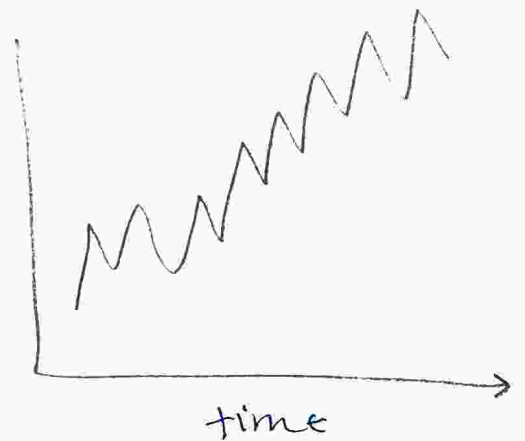
→ linear model

$$T_t = m \cdot t + b$$

→ detrended series

$$Y_t^1 = Y_t - T_t$$

(trend component) السلسلة
(time series) السلسلة



Differencing

↳ technique used to remove non-stationarity ~~(technique)~~ ~

$$x(t) - x(t-1) = ARMA(p, q)$$

→ done by subtract observation in current period from previous one.

⇒ ARIMA Parameters

p: AR & d: I & q: MA

Where

p: no. of autoregressive terms.

d: no. of differences.

q: no. of moving average terms.

(notes)

Auto correlations

↳ numerical values that indicate how a data series is related to itself over time.

↳ measures how strongly data values at a specified number of periods apart are correlated to each other over time.

(9)

* Auto Correlation Function (ACF)

↳ Correlation of the values of time series with itself.

↳ Helps to determine the order, q of MA model.

* Partial auto correlation function (PACF)

↳ AutoCorrelation calculated after removing the linear dependence of previous terms.

↳ Helps to determine order, p of AR model

AR \rightarrow Auto-regression

MA \rightarrow moving average

* How to do a time series analysis

1) visualize time series

1) visualize time
(data) (explore) له انك تعلم
(model) له صغر ال (complexity)

لم مصغیر ال (exploration) مش متعرف ال (series) تک نه
(stationary) ولا لا.

2) stationarize the series

8 detrending not allowed. ~~not allowed~~

- Seasonal adjustment & differencing

3] Plot ACF/PACF to find optimal Parameters

له ال Parameters (P, d, q) مكر تيجر باستفهام ACF, PACF
له لو ACF, PACF بقل ~~statistical~~ تدريجي، معناه
لنا فعل ال (series) ← stationary.

4] Build ARIMA model

← مستفهم ال Parameters (model) عناه نيج ال

5] Make Predictions

→ له حالياً قدر فعل (Predictions).

Time series

Reasons to choose

Minimal data collection.
↳ don't need to input drivers

designed to handle the
inherent^{ent} autocorrelation
of lagged time series.

Accounts for trends
and seasonality.

Cautions.

no meaningful drivers.
↳ no explanatory value.
↳ can't stress test.

↳ It's an "art form" to
select appropriate
Parameters.

→ only suitable for
Short term Predictions.

Lec 14

text analysis

↳ processing & representation of text for analysis and learning tasks.

* Main challenges in text analysis:

1) High-dimensionality

↳ every distinct term is dimension.

↳

2) data is un-structured

* Inverse document Frequency (idf)

$$\text{idf}(t) = \log_2 [N / \text{df}(t)]$$

↳ indicates the importance of:

↳ search (relevance)

↳ classification

مع الحاضرة محتاج تتذكر مدار (slide) أكثر

لأن جانب يفتحها كلها.